# DAT 375 Project One Template

# Data Analysis Process Job Aid

# Daniel Gorelkin

This report is meant to be used as a reference guide for new staff members of SNHU Consulting and to act as a job aid for creating a Storm and Crime Data Report (SCDR) as requested by the City of Miami Police Department.

### Introduction

The SCDR report should provide the police department with information they can use to anticipate possible timeframes and possible crimes in the future in the City of Miami, based on the provided historical data from October 2019. The report will include descriptive statistics, correlation analysis, and linear and/or logistic regression analysis to assist the police department in making the correct decision based on variables such as date and time, crime type, storm type as categorical variables, and the city zone or crime area.

### Type of analysis

The discussed problem in the SCDR report should provide a model that will be able to predict possible crime events as an outcome of stormy weather events. To do so, we should first check and validate whether the connection between crime and storms indeed exists, as the police department assumes, and find patterns within the given timeline period. However, we should bear in mind that some potential barriers might exist. Hence, we should check for completeness of the data in the dataset, e.g., all fields for the date, crime, storm event, and the crime location must exist and contain no null or blank records. In the same way, we should filter the fields where a crime event occurred without having a storm and vice versa. This means that our data should be cleaned and sorted before conducting the analysis. Additionally, we should check that the crime type and storm events are grouped and categorized. We should also not forget about the volume of data, which must be large enough to conduct the training and testing for the prediction regression model and comply with statistical standards.

Hence, to complete the SCDR report, we will need to conduct a series of analyses, such as descriptive statistics, to summarize the dataset and provide a visual and quantitative representation of the given events. This will assist us in communicating the findings to non-technical decision makers in the police department. Similarly, to test if indeed a connection between crime and weather exists, we will run a correlation analysis using Pearson correlation and correlation matrices. This will help us to conclude whether a correlation between storms and crime exists and its direction. Ultimately, we will build a linear regression analysis model that will establish the relationship between any of the weather variables, such as the area in the city or the type of storm. Additionally, we could build a logistic regression model that can provide a qualitative outcome of whether crime will occur. By doing so, the models will determine the likelihood of a potential crime in the inspected area and will assist the police in making the right call.

### Defining Parameters and Collecting the Data

To determine the relevant data variables to conduct the analysis of the given problem, we will need to fetch and group the weather variables, such as the type of storm event and the date it occurred. For example, we can assume that we will have various storm events that triggered crimes on several distinct dates. Hence, we will need to differentiate the effects of each type of storm event and define them as qualitative categorical variables, as shown in the figure below:

A screenshot of a computer

AI-generated content may be incorrect.A computer code with text

AI-generated content may be incorrect.

The command output above will expose the various storm activity types that could lead to a crime event.

Similarly, we should determine the criminal data variables required for this analysis, such as the type of criminal activity, violence, theft, or robbery, and define them as categorical variables as well. Again, we should pay close attention to the date of the event and make sure that it occurred while a storm event took place. Additionally, we might want to know the criminal frequency, such as how many times each crime is repeated within the examined timeline. For instance:

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer code

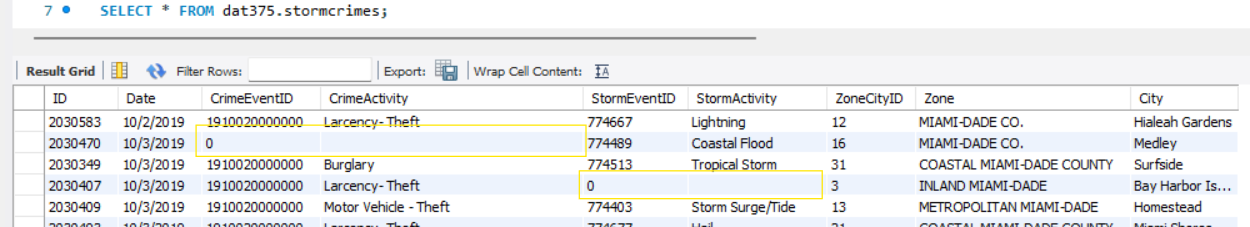
AI-generated content may be incorrect.

Finally, in addition to the date and time when the storm or crime occurred, we will need data about the geographical zone in the Miami district where the crime took place. Therefore, to conduct a thorough and insightful investigation, we should aggregate data by date, crime type (categorical), storm type (Categorical), and the city zone or area. It will also be useful to collect all available data for each single day of the inspected month, even if such a day had no crime or storm event, to keep track of how often storms occurred and how many crimes were committed during storms versus non-stormy days:

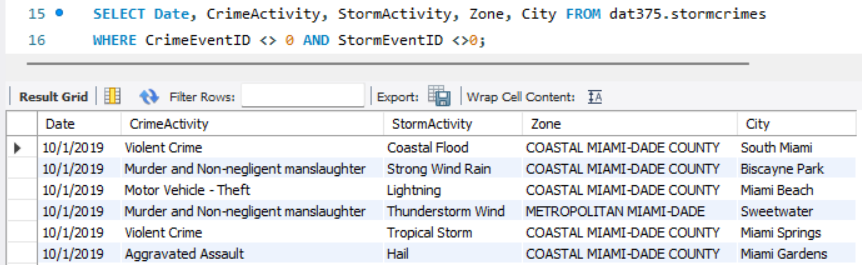
A screenshot of a computer

AI-generated content may be incorrect.

### Appropriate Tool Selection

With that in mind, we will use several statistical tools to produce the report. To traverse and explore our dataset, we might pull SQL data directly from the database, which will provide us with a greater understanding of the records and will help us to identify messy data. For example, the following query will give us an idea that there are some missing records in the dataset that should be cleaned before processing the data:  


Similarly, it will show us that some events of crime could occur while there is no storm event in place, and the other way. Hence, to filter the events where no crime or no storm event took place, we can run:



After getting familiar with our data, we will switch to our next analytical tool such as Python, in which we can use *SciPy, Pandas,* and *NumPy* to clean the data and summarize the analysis. Similarly, we can consider using SAS or R language, which are more suitable for and require less coding time and effort. Despite that SAS offers plenty of powerful packages to conduct all types of statistical analysis we need, and because SAS is licensed software which our office don’t have the access to, we will conduct the analysis in RStudio via the R language, which is an open source and very powerful tool that can both clean and analyze the data and provide the appropriate visualizations to demonstrate and interpret the outcome of the analysis. For example, we can download and use the *tidyverse* library in R by running *install.packages("tidyverse")*, and run library*(tidyverse)* to load its core packages, e.g., *tidyr* for cleaning data, *lubridate* for handling working with date-time variables, and *ggplot2* for creating visualizations.

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

However, despite the capabilities of using R for insight visualizations are great, I might recommend using another visualization tool, such as Power BI, which can read the exported data from R and visualize the analysis so that it will be clear and easy to read for the stakeholders. For example, the following output presents the spread of crime types in the Miami area while considering the type of storm on that date:

A graph of different colored lines

AI-generated content may be incorrect.

### Section 4: Validation

To validate that we have the correct data parameters by sampling our dataset, besides querying the information directly from the dataset, as [displayed above](#data_from_a_larger_dataset) by using SQL, we can run scripts to filter the needed data in R (or Python or SAS) as well. For example, we can remove the unwanted columns from the dataset and ignore the rows that don’t represent a crime conducted on a stormy day:

# Remove irrelevant columns

*data\_reduced <- data %>% select(-CrimeEventID,-StormEventID,-ZoneCityID)*

# Clean and preprocess the data

storm\_crime <- data\_reduced %>%

mutate(Date = mdy(Date), CrimeActivity = str\_trim(CrimeActivity), StormActivity = str\_trim(StormActivity),) %>% filter(!is.na(Date), !is.na(CrimeActivity), !is.na(StormActivity))

# Factor the needed variables

storm\_crime\_data <- within(storm\_crime, {CrimeActivity <- factor(CrimeActivity), StormActivity < factor(StormActivity), Zone <- factor(Zone), City <- factor(City)})

# Display a sample of the cleaned and filtered data:

A computer screen shot of text

AI-generated content may be incorrect.

# Output cleaned data to a fresh CSV file to be passed to the next platform and stakeholders' validation



Similarly, we can look into plots and get visual insights, for example, from the plot of the volume of crime in the Miami area during the stormy days of October 2019, with a breakdown of crime type (*CrimeActivity*). From this output, we can clearly see that Coastal Miami-Dade County faced increased crime activity compared to all other zones during storms.

A graph of different colored bars

AI-generated content may be incorrect.

Lastly, before deploying the models and summarizing the report, we shall not forget to follow the six phases of CRISP-DM and validate the selected parameters and the collected data with our decision-makers. We must present to them the main parameters and their functionality in the model, such as *CrimeActivity*, *StormActivity*, and the *Zone,* and *City* variables. We must validate that indeed the output contains all types of interest and make sure if any crime types are missing or irrelevant to the model and modify it, or, if additional filters and drilling in is desired to achieve a more focused output.

### References

*Clarke, E. (2024, June 12). SAS vs R: What is Difference Between R and SAS? Guru99.* [*https://www.guru99.com/sas-versus-r.html*](https://www.guru99.com/sas-versus-r.html)

*Larose, D. T. (2015-04-13). Data Mining and Predictive Analytics, 2nd Edition. [[VitalSource Bookshelf version]]. Retrieved from vbk://9781118991121*

*Luna, J. C. (2022, December 28). Python vs R for data science: Which should you learn?. DataCamp. https://www.datacamp.com/blog/python-vs-r-for-data-science-whats-the-difference*

*Power BI - Data Visualization | Microsoft Power Platform. (n.d.).* [*https://www.microsoft.com/en-us/power-platform/products/power-bi*](https://www.microsoft.com/en-us/power-platform/products/power-bi)

*Tidyverse packages. (n.d.).* [*https://www.tidyverse.org/packages/*](https://www.tidyverse.org/packages/)